IS-ENES3 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°824084

IS-ENES3 Virtual First General Assembly

SESSION 5 – Cross-WP issues

Models, tools & HPC

Chat & discussion supporting document

Friday 27 March, 2020

10h-13h CET

This document, editable by any participants, is to help moderate questions and support discussions during the Virtual General Assembly.

During presentations, please write down your full name and your questions/comments in this document, below the dedicated talk. This process allows to keep track of questions, avoid their repetition and foster discussion.

Cross-Wp issues

10h-11h30 XIOS benchmarks - *Sophie Valcke, Mario Acosta*

- ○ Meeting URL: https://cerfacs.webex.com/cerfacs-en/j.php?MTID=m2cdada7feb5d88161241ecd0036d273d

Name of participants:

Marie-Pierre Moine, Grenville Lister, Jean-Christophe Rioual, Bryan Lawrence, Gaelle Rigoudy, Xavier Yepes, Yann Meurdesoif, Uwe Fladrich, Kim Serradell, Stella Paronuzzi, Italo Epicoco, Massimiliano Drudi, Sylvie Joussaume, Olivier Marti, Miguel Castrillo

Summary / main actions / important points **(people in bold are responsible to take the action forward):**

- Organise a telco to discuss use of ESDM in XIOS (**Bryan**) ; participants: Julian Kunkel, Yann Meurdesoif, Bryan Lawrence, Sophie Valcke, Mario Acosta, Italo Epicoco, Xavier Yepes, Uwe Fladrich, Miguel Castrillo, Kim Serradell, Grenville Lister.
- XIOS benchmark developments are part of ESiWACE2 and should be reported there (it was however OK to discuss the strategy around XIOS benchmarks as part of IS-ENES3 networking activities)
- Remarks on the benchmarks presented:
  - Attached mode (i.e. no XIOS server; it is the XIOS client that writes the file) should not be used in the benchmarks as it has not been designed for performance, just for facilitating development.

  - The poor performance of one_file mode should be investigated further; profiling of the performance of multiple_file mode on MN4 should be done to compare (related issue: use of ESDM to solve this issue?)
  - Benchmarks should include performance analysis of XIOX filters
- Proposition to organize a monthly telco around XIOS to share experience between people working with XIOS (not for support) (**Sophie**); examples of issues discussed in this BOG that would/will be relevant for these monthly telcos :
  - Issue of offloading the filters from the client to the server side; possibility of using MPI3 shared-memory for intra-node offload and asynchronous treatment.
  - performance of spatial filtering
  - use of XIOS2.5 with NEMO ORCA36
  - uniformisation of XIOS and OASIS performance timers and load-balancing analysis tools (like LUCIA) for analysing/optimising ESMs (low hanging fruits: same write out format)
  - tool to help the user determine a good XIOS configuration, exchange heuristics on how to tune XIOS, step-by-step guide on how to use XIOS?

  What should be the media used to keep a trace of these discussions? Do we need to set-up an XIOS forum?
- XIOS memory consumption is currently a problem. Monitoring of memory consumption should be included. Yann has ideas on how to solve this but it is a big task, also linked to the workflow reorganization.
- Yann is working on a new version including a reorganization of XIOS workflow for better performance. The development of this new version should not slow down the development of benchmarks as using the benchmarks on both the current and the new version to evaluate the gain will be interesting.
- Need for an XIOS roadmap: an XIOS advisory board has been set up and a development plan should be discussed soon (**Yann, Sophie**)

Discussions during the BOG

Name: type your question/comment

- Italo Epicoco: when mention “online postprocessing” do you refer to the online diagnostics or to operations at the end of the simulation after the model run is completed?
  - Bryan: I presume these are using the XIOS filters? Yes
- Grenville: what is the wallclock time taken for the tests?
- Bryan: I think it should be possible to start using the ESDM in these tests, so you can use multiple filesystems on write … (and hence more OSTs)
  - (We can talk about this at a special telco if desired)
  - In principle the ESDM should handle some of the issues around number of OSTS and files. The ESDM is now implementable by using a revised NetCDF

IS-ENES3 has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N°824084

library, and provided we sit within the "supported NetCDF operations", it should help with some of the POSIX related issues.

        - (Who) could take part in a telco with Julian Kunkel to discuss the possibility of doing such a benchmark?
        - (Can people who would want to take part in such a telco add their names here:
            - Bryan Lawrence
            - Sophie Valcke
            - Mario Acosta
            - Italo Epicoco
            - Xavier Yepes
            - Uwe Fladrich
            - Miguel Castrillo
            - Kim Serradell
            - Grenville Lister

- Marie-Pierre : In your plots on scalability, y-axis is total elapse time or cpu time?
    - Answer: total time from beginning to the end of the simulation
- Marie-Pierre: numbers for one_file mode you show : is it with the 2nd level of XIOS servers activated ?
    - Answer: no because using XIOS 2.0 (functionality not in)
- Uwe Fladrich: Are the grids Tco? Because you refer to CMIP6 AMIP/HighResMIP, which use TL grids.
    - Answer: it is Tco
- Italo Epicoco: in your plots of scalability, the total number of cores is kept constant and only changes the number of XIOS servers?
    - Answer: yes, the number of cores for the model does not change
- Jean-Christophe: For our current generation of models (Met office - CMIP6), the XIOS cost is not IO offloading but temporal filtering on client side (daily or monthly average of timesteps). It would be useful to benchmark these too, not just instantaneous values at high frequencies.
- Sylvie Joussaume: glad IS-ENES3 GA can host this community discussion ! However, you will have just to be clear where the reporting of the activity will be: it seems more relevant to ESIWACE2. Role if IS-ENES3 to trigger community discussion is very good. Implications of this work on IS-ENES3 has also to be clarified. So excellent to have this discussion now !
    - Kim: I think you're highlighting an important point here, Sylvie. BSC plan was to report his activity in ESiWACE2. But results from these benchmarks can have an important impact directing the next developments for XIOS in IS-ENES3.
    - Sylvie: so we agree ! fine

- Uwe Fladrich: Regarding the affinity: Is it not impossible to place the IO servers "near" the computation? Because this would mean, ultimately (i.e. at scale), to place an IO server with each compute node. Cf. also Yanns comments on the attached mode.
- **Yann :** Attached mode has not been designed for performance, just for facilitating development. Why to use it in such benchmark ?
- Jean-Christophe: Met Office is interested in the performance of spatial filtering between different grids (unstructured -> structured). Has anyone any experience to share ? Both computational performance and numerical evaluation of the interpolation.
- **Miguel:** Italo, I guess you were always using XIOS 2.0 in your test right? We had lots of problems running ORCA36 with XIOS2.5.
    - Italo Epicoco: we used XIOS2.5 but only developing an ad-hoc client not running the ORCA36 model, but allocating fields of the same order of those of ORCA36 resolution
    - Miguel: Thank you. True, I remember you mentioned it at the beginning. In that case we are interested in knowing if ORCA36 should be expected to work on XIOS2.5 (or more if further developments will be built from XIOS2 or XIOS2.5 (CMIP6-only version?), or maybe from none of these). As far as I know now one is testing this configuration with X2.5, only 2.0.
- UF: Regarding the XIOS profiling and help for users: Could there be a connection to the load-balancing problem and tools (like LUCIA) for analysing/optimising this for ESMs? Any coordinated development possible or at least definition of metrics and presentation of results?
    - Mario Acosta: I agree that it could be a connection and it would be a nice idea. We are exploring the load-balance problem and we can write suggestions for this connection with XIOS part, but as you know it is a complex process.
    - Pick low hanging fruits. For example, have any tool write out performance data/hints in a common, easily parsable format. Make this consistent, for example, across XIOS and LUCIA.

- Grenville; Would it be possible to develop a tool to help the user determine a good XIOS configuration.

  Difficult to do in general - XIOS logs do provide hints.

  Jean-Christophe: may be could exchange heuristics on how to tune XIOS for users that have working knowledge of XIOS but are not necessarily HPC optimisation experts.

Main actions decided:

Possibilities before final writing:

- BSC benchmark proposed (comments/suggestions)
    - Profiling of the poor performance of one_file mode
    - Profiling of performance of multiple_file mode on MN4 to compare
    - Include filtering tests
- ESDM test?

  - Kim: BSC will contact Julian K. to coordinate a test with ESDM.
- Utility to set up optimal configuration for XIOS?
- Ensure that memory consumption is not a problem for future versions
- New tests after XIOS Team modifications?
- Monthly telcos?
- Roadmap and groups creation

11h30-13h Computational evaluation of ESMs, including coupling issues (LUCIA) and energy consumption - *Mario Acosta*

  - Meeting URL: https://cerfacs.webex.com/cerfacs-en/j.php?MTID=m0a38658359a59de245bb2c99c3712fe9

Name of participants: Sophie Valcke, Italo Epicoco, Uwe Fladrich, Sergi Palomas, Xavier Yepes, Miguel Castrillo, Grenville Lister, Harry Shepherd, Florian Ziemen, Gaëlle Rigoudy, Alok Kumar Gupta, Marie-Pierre Moine

Summary / main actions / important points (people in bold are responsible to take the action forward):

Discussions during the BOG

Name: type your question/comment

- Sophie Valcke: I don't understand "An alternative could be to approximate this value from the Linpack execution "
- Sophie Valcke: Anyone interested in collaborating with Mario and Italo on getting the energy counters that were just described (HPM counters, RAPL and NVML libraries, etc.)?
- Italo Epicoco: instead of SYPD we should use the execution time for load balancing
- Uwe Fladrich: … or the (minimised) waiting time
- Miguel Castrillo: CHPSY (core hours per simulated year) is a secondary energy measure that can be used here.
  - Italo Epicoco: I agree. the CHPSY can be used when the "cost" of resources matters, while we can use the time to solution when the "cost" is less important. However, the aim of load balancing remains the same: don't waste resources neither time.
- Harry Shepherd: Given the variations of the model execution times on HPC, we need to choose our load balancing such that at no point is the model that uses the greatest resource is waiting

- Jean-Christophe: A remark. In practice, for CMIP6 production, we tend to fix SYPD and try to minimise node count rather than minimise elapsed time.
    - Miguel: I think this is valid for an energy-to-solution approach, that should be the standard. However, sometimes deadlines impose a different schedule and a more time-to-solution approach. Additionally, because of shared-resources contention/re-use (memory, cache) sometimes the more efficient configuration needs to use a minimum amount of resources and thus may achieve a minimum SYPD. It may seem counterintuitive but with some configurations you cannot achieve so much efficiency with let's say 16 nodes than with 32.
- Jean-Christophe: It should be easy to normalise the output by core count to produce a load balancing plot. It is easy for user to interpret.
- Harry Shepherd has uploaded to the Slack Channel an example of the load balancing plots used by the Met Office, that we have had used successfully to help load balance coupled models.
-